Appendix to Subcommittee Report

This document is a portion of an approximately 75-page summary of review materials distilled from (1) around 10,000 pages of application materials by accrediting agencies to the Department of Education for recognition, and (2) publicly available sources on assessment of student learning and its relationship to accreditation compliance. The summary document was compiled by the subcommittee chair (David Eubanks) as preparatory material for interviews and discussions related to the subcommittee's charge.

This appendix contains only portions of the summary document drawn from public sources, mostly academic journals, that offer different perspectives on the complexities of measuring student learning in the context of accreditation compliance. References to specific accreditors and non-public materials have been removed.

The materials here should be seen as informal (not peer-reviewed) and not comprehensive. This appendix is not endorsed by the subcommittee as part of the report itself but is provided as background material for those wishing to research these topics themselves.

Scholarship on Learning Assessment

Overview of Opinions in Higher Education

A brief history of the assessment movement is given in Bass, et al (2016) in the context of a qualitative study of public comments on the topic. From the abstract: "[I]t is no secret that efforts to foster a 'culture of assessment' among institutions of higher learning have frequently encountered resistance, particularly on the part of faculty unconvinced that the aspirations of the assessment movement are in fact achievable."

The article contrasts assessment fans and detractors by classifying comments made on public forums, and can serve as an introduction to the topic. The discussion is framed with these contrasting epigrams:

> Outcomes-assessment practices in higher education are grotesque, unintentional parodies of both social science and "accountability." No matter how much they purport to be about "standards" or "student needs," they are in fact scams run by bloodless bureaucrats who do not understand the holistic nature of a good college education.
>
> —Fendrich (2007)

> Those who are afraid of rubrics and assessment instruments remind me of Luddites who refuse to perceive reality. If we are to rely on our time-tested bold statements that "we are a quality institution," without any evidence, then we deserve to be judged by outside constituencies.
>
> —Anonymous

The authors explore these viewpoints as opposites, but there is another interpretation—that both sides can be correct. After all, assessment has been an integral part of teaching for a long time (giving exams, grading papers, etc.). The subsections below lead to a possible explanation for the vigor of the resistance to "assessment" as associated with accreditation: that the expectations for peer review are not as useful to faculty as the assessments they get to choose (and use) themselves.

Reference:

Baas, L., Rhoads, J. C., & Thomas, D. B. (2016). Are quests for a "culture of assessment" mired in a "culture war" over assessment? A Q-methodological inquiry. *Sage Open*, 6(1), 2158244015623591.

Accreditation Constraints

Blaich & Wise (2018) describe the constraints of scope, cost, and speed in preparing assessment reports for accreditation and how those affect results. They note the tendency toward *pro-forma* reports (i.e. cookie-cutter reporting) due to the stress of accreditation (pg. 75):

> In our experience, getting a warning about assessment from an accreditor tends to focus the minds of administrators on the importance of assessment–or at least on the importance of not getting "dinged" by an accreditor. Too often, such warnings lead to an all-hands-on-deck response—one which trades speed for quality. As the Iron Triangle tells us, if you try to complete big projects on a short timeline with limited resources, quality inevitably suffers. And worse, externally-driven, breakneck efforts to "fix" assessment programs can reinforce the idea among faculty and staff that assessment is nothing more than make-work imposed by outsiders who are hostile to their work.

For example, they note the common outcome of "process churn" (changing process instead of producing results), pg. 75:

> One can close the loop endlessly by changing assessment measures, changing the criterion for demonstrating competency, changing the outcomes, or changing the way that student work is sampled—assessment full of sound and fury, signifying nothing.

On scope, they take issue with the normal procedure of continual reporting on several learning goals per year (pg. 76)

> We encounter too many places where faculty and staff are working on closing the loop for three or four outcomes a year. How can someone possibly determine whether a change at the program or department level has resulted in improved student learning in one year?

They conclude with (pg 76):

> We do not blame either the economy or accreditors for what's happening. What we do assert is that we need to acknowledge the corner we're in when it comes to assessment and recognize that unless we change the scope, cost, and speed equation, we are inadvertently promoting processes that are better at generating reports than improving student learning.

Looking for Results

In the research literature, the empirical benefits of a standardized assessment program are hard to find, and the methods are questioned.

> Although assessment is widespread, examples of assessment data, including survey results, being used to effect institutional change are relatively rare (Banta, 2002; Ewell, 2002; Peterson et al.). Pike (2002, p. 147) concluded "there is no greater problem in assessment than our inability to influence decision making with assessment results."

This excerpt comes from

> Pike, G. R. (2006). The convergent and discriminant validity of NSSE scalelet scores. *Journal of College Student Development*, *47*(5), 550-563.

The references from the quote above are:

> Ewell, P. T. (2002). An emerging scholarship: A brief history of assessment. In T. Banta (Ed.), *Building a scholarship of assessment* (pp. 3-25). San Francisco: Jossey-Bass.
>
> Peterson, M. W., Einarson, M. K., Augustine, C. H., & Vaughan, D. S. (1999*). Institutional support for student assessment: Methodology and results of a national survey*. Stanford, CA: National Center for Postsecondary Improvement.
>
> Pike, G. R. (2002). Measurement issues in outcomes assessment. In T. Banta (Ed.), *Building a scholarship of assessment* (pp. 131-147). San Francisco: Jossey-Bass.

A more recent article is

> Banta, T. W., & Blaich, C. (2010). Closing the assessment loop. *Change: The Magazine of Higher Learning*, 43(1), 22-27.

These authors wrote that "We scoured current literature, consulted experienced colleagues, and reviewed our own experiences, but we could identify only a handful of examples of the use of assessment findings in stimulating improvements."

Trudy Banta is one of the founders of the assessment movement, an author of the AAHE Principle of Assessment, and led the Assessment Institute at IUPUI for years.

Problems with Methods

One analysis of assessment methods centers on the nature of evidence and the trade-off between quality and quantity:

> The dysfunctionality of assessment today starts with the primacy of evidence and data. One of the key premises of the assessment paradigm is that the faculty's conventional wisdom about what students can and cannot do well is unreliable. We therefore must collect direct evidence of

> students' abilities to master the outcomes that we define to be part of their educational process. […] The problem is that assessment data can be either cheap or good, but are rarely both. The fact is that good evidence about student learning is costly.

A more specific criticism is made against the practice of generating data by comparing student work to a rubric (a scoring guide) to assign a level of proficiency.

> Unfortunately, creating good data from a rubric-scoring process is very difficult—and the availability of substantial resources makes it only slightly less so. The main problem is that scoring is a necessarily subjective process that requires all kinds of judgments about what key terms mean, how to distinguish between performance categories, and how to sort students' work into those categories. Calibration sessions that give readers training can be helpful, at least in promoting greater reliability, but may not help in establishing validity (e.g., everyone agrees what a "proficient" response looks like for scoring purposes, but is a student who wrote it really and truly *proficient?*). Moreover, it is often difficult to determine whether a student is unable to show mastery of an outcome or whether an assignment just didn't do enough to prompt the student to show that mastery.

Reference: Roscoe, D. D. (2017). Toward an improvement paradigm for academic quality. *Liberal Education*, 103*(1)*. Also see Linda Suskie's (assessment consultant and author) blog post praising this article[1], which also notes that "[W]e've left a vital part of the higher education experience—the grading process—in the dust. We invest more time in calibrating rubrics for assessing institutional learning outcomes, for example, than we do in calibrating grades."

The necessity to gather "cheap" data can result in lower quality, as noted in the quote, but also lower quantities, which exacerbates statistical problems.

A critique of sample sizes typically used in assessment is found in

> Bacon, D. R., & Stewart, K. A. (2017). Why assessment will never work at many business schools: A call for better utilization of pedagogical research. *Journal of Management Education*, 41(2), 181-200.

These authors find that in using plausible assumptions about measurement error, the amount of data required to make decisions is larger than what is typically available.

A more recent article in the same vein is

> Fulcher, K. H. & Prendergast, C. O. (2019) Lots of Assessment, Little Improvement? In Hundley, S. P., & Kahn, S. (Eds.). (2019). Trends in Assessment: Ideas, Opportunities, and Issues for Higher Education. Stylus Publishing, LLC.

While educational research on learning is thriving (there are several journals just on writing assessment, for example, and see *Educational Data Mining*), the focus of assessment reporting for accreditation seems to be on process, not empiricism. This communication gap is unfortunate, because there is much

[1] https://www.lindasuskie.com/apps/blog/show/44545247-a-new-paradigm-for-assessment?siteId=115520809&locale=en-US

research that could inform assessment practice for program improvement. For example, the difficulties in measuring writing (and the correlation with grades), for example in

> White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. University Press of Colorado.

Critical thinking is a common learning goal in use, but the efficacy of improvement programming is questionable, as reviewed in

> Huber, C. R., & Kuncel, N. R. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research*, *86*(2), 431-468.

The authors conclude that "[A]nalysis of curriculum-wide efforts to improve critical thinking indicates that they do not necessarily produce incremental long-term gains."

Another study with practical consequences to program assessment is a study from AAC&U:

> Sullivan, D. F., & McConnell, K. D. (2018). It's the Assignments—A Ubiquitous and Inexpensive Strategy to Significantly Improve Higher-Order Learning. *Change: The Magazine of Higher Learning*, 50(5), 16-23.

The authors conclude that simply averaging VALUE rubric score can lead us to believe that no learning has taken place when in fact it has (one needs to consider inputs). How many improvement programs have been based on misinterpreting VALUE rubric scores because of oversimplified analysis?

Other references:

Hathcoat, J. D., Penn, J. D., Barnes, L. L., & Comer, J. C. (2016). A second dystopia in education: Validity issues in authentic assessment practices. Research in Higher Education, 57(7), 892-912.

Finding Improvements

The learning assessment standards do not address how programs are supposed to use data to find improvements. It does not follow that knowing a state of affairs (e.g. blood pressure, stock market prices) entails knowing how to act to achieve goals. With the unit report system of addressing learning, programs seem to be expected to discover interventions on their own. This problem is addressed in a recent article (Finney & Buchanan, 2021, pg. 37):

> There is great inefficiency in the outcomes assessment process when programming is either built from "scratch" based on good intentions, assumptions, and hunches, or programming is based on ineffective strategies. Depending on the initial quality of programming, major changes may be required for programming to be effective. Although outcomes data can indicate that students did not achieve expectations, outcomes data do not suggest changes to programming. Moreover, faculty and student affairs professionals may not know what programming is necessary to achieve intended outcomes (e.g., Brownell & Tanner, 2012; Hutchings, 2010; Jones, 2008). Thus, changes to programming may be exploratory in nature (e.g., "Let's try this approach"), based on tradition (e.g., "This is what I experienced as a student"), or avoided. Moreover, changes may be minor. Thus, it may take years to implement effective programming

> that results in intended outcomes. An analogy offered by Eubanks (2017) makes this point clearly: "Imagine if each town and village were required to research and produce its own drugs, and ignore large-scale science-based medical research. That is our current situation with respect to assessment" (p. 11).

The unit-reporting expectations for learning standards work against a research-based approach to improvements by requiring so many solutions to unique problems (typically one per program). Research-based projects may be most effective when they affect the most number of students, likely by crossing program delineations. As an example, consider the findings from a credible study at the U.S. Naval Academy (Insler, et al, 2021, pg. 1):

> Exploring a variety of mechanisms, we suggest that instructors harm students not by "teaching to the test," but rather by producing misleading signals regarding the difficulty of the subject and the "soft skills" needed for college success. This effect is stronger in non-STEM fields, among female students, and among extroverted students. Faculty that are well-liked by students—and thus likely prized by university administrators—and considered to be easy have particularly pernicious effects on subsequent student performance.

If this finding is generally applicable, then adjusting faculty feedback to improve consistency and rigor of grading would lead to broad learning gains. This scenario does not fit within the unit-reporting format, however, since it would have to identify learning goals one by one first, measure them, then justify action, and so on.

In Blaich & Wise (2011), the authors note that institutions often have plenty of data, but there are obstacles to using it (pg. 12), echoing Machiavelli's observation about the difficulty of changing the order of things.

> Even when assessment reports are disseminated widely, most of us behave as though the data in the reports will speak loudly enough to prompt action. We tend to believe that interesting findings will naturally prompt discussions and ultimately revisions in our courses and programs. But this denies the reality on most of our campuses—that the current state of affairs in our departments, curricular structures, and programs is usually a compromise carefully negotiated among numerous parties over the course of years. Unless the findings are truly devastating, assessment data has little impact on this tightly constrained arrangement.

Their recommendation is to focus efforts on a single project and ensure buy-in and resources from the beginning, rather than many diffuse projects.

References:

Blaich, C. F., & Wise, K. S. (2011, January). From gathering to using assessment results: Lessons from the Wabash National Study. (Occasional Paper No. 8). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA)

Blaich, C. F. & Wise, K. S. (2018) Scope, Cost, or Speed: Choose Two—The Iron Triangle of Assessment, *Change: The Magazine of Higher Learning*, 50:3-4, 73-77, DOI: 10.1080/00091383.2018.1509606

Finney, S., Buchanan, H. (2021) A more efficient path to learning Improvement: Using repositories of effectiveness studies to guide evidence-informed programming, *Research & Practice in Assessment,* (16)*1, 36-48*

Insler, M., McQuoid, A. F., Rahman, A., & Smith, K. A. (2021). Fear and Loathing in the Classroom: Why Does Teacher Quality Matter?. *IZA Institute of Labor Economics.* DP 14036.

References Cited in the quotes:

Brownell, S., & Tanner, K. (2012). Barriers to faculty pedagogical change: Lack of training, time, incentives, and... tensions with professional identity? *CBE—Life Sciences Education*, 11, 339 – 346. 10.1187/cbe.12-09-0163

Eubanks, D. (2017). A guide for the perplexed. *Intersection: A Publication of the Association for the Assessment of Learning in Higher Education*, Fall, 4-13.

Hutchings, P. (2010). Opening doors to faculty involvement in assessment. (Occasional Paper No. 4). University of Illinois and Indiana University, *National Institute for Learning Outcomes Assessment*.

Jones, A. (2008). Preparing new faculty members for their teaching role. *New Directions For Higher Education*, 143, 93-100. https://doi.org/10.1002/he.317

Need for Innovation

The need to improve learning assessment processes was noted in detail in a "grand challenges" call published by NILOA:

> Singer-Freeman, K., & Robinson, C. (2020, November). Grand challenges in assessment: Collective issues in need of solutions (Occasional Paper No. 47). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

The need for the grand challenges is associated with a lack of credibility of current efforts (pg. 4).

> Perceptions about the value and worth of assessment in higher education at times are bleak. A survey of chief academic officers found that almost a third thought assessment was for appeasing politicians and accreditors as opposed to improving teaching and learning, while almost a fifth did not agree that systems of assessment have led to improvements in quality of teaching and learning (Jaschik & Lederman, 2020). This is not a new narrative to assessment professionals who are well aware of the negative perceptions bestowed upon assessment, and frequently struggle to assuage negative perceptions of assessment (Ariovich et al., 2019). However, as is often the case with wicked and complex issues, simply ignoring them will not make them go away.

A few of the grand challenges are (quoting from the source):

- Innovation. "The process of assessment should produce visible and actionable assessment findings that drive innovation." (pg. 5)

- Budgeting. "Assessment findings should be used to inform budgetary decisions." (pg. 5)

- Quicker improvements. (pg. 6):

  Too often, assessment findings are not utilized to direct immediate pedagogical improvements, in part because the work of closing the loop in student learning outcomes assessment is too slow to benefit the students who are assessed or to improve the instruction or course design of those who are teaching (Eubanks, 2017; Maki, 2017). We must find ways to make changes in response to assessment findings within the space of a single class through formative assessments (Dirlam, 2017; López-Pastor & Sicilia-Camacho, 2017; Maki, 2017).

- Equity of outcomes. (pg. 6):

  Assessment findings should be used to increase educational equity. A goal in higher education is that every student has an equal opportunity to succeed regardless of ethnicity, gender, socioeconomic status, ability, or family educational history. There is compelling evidence that we are not meeting this goal (Cahalan et al., 2018).

- Disaggregation. (pg. 7):

  Data on learning should be disaggregated to consider important student characteristics. For the most part, assessment data are reported and reviewed in ways that mask inequities because student learning outcome reports aggregate assessment results across all sections of courses and instructors.

- Change over time. (pg. 7) "To identify progress, it is essential to examine changes in institutional effectiveness (including student learning) over time."

The fact that these basic issues are still challenges after decades of accreditation assessment reporting implies that the system has not worked very well and/or is not improving itself. The authors sum up the need for change with (pg. 9)

> Assessment in higher education has a range of challenges related to improving measurement, addressing inequities, and fostering continuous improvement. To enable collective solutions to these grand challenges, the shift away from compliance towards assessment for improvement needs to be fully realized.

References cited:

Ariovich, L., Bral, C., Gregg, P. L., Gulliford, M., & Morrow, J. A. (2019). The assessment profession in higher education: A snapshot of perceptions, roles, and activities. *Assessment Update*, 31, 10-12. doi:10.1002/au.30175

Cahalan, M., Perna, L. W., Yamashita, M., Wright, J. & Santillan, S. (2018). 2018 indicators of higher education equity in the United States: Historical trend report. Washington, DC: The Pell Institute for the Study of Opportunity in Higher Education, Council for Opportunity in Education (COE), and Alliance for Higher Education and Democracy of the University of Pennsylvania (PennAHEAD).

Dirlam, D. K. (2017). Transformative learning needed for higher education assessment. Emerging Dialogs in Assessment. Retrieved from: https://www.aalhe.org/page/ed_2017transformativ/EMERGING-DIALOGUES-IN-ASSESSMENT.htm

Eubanks, D. (2017). A guide for the perplexed. *Intersection: A Publication of the Association for the Assessment of Learning in Higher Education*, Fall, 4-13.

Jaschik, S., & Lederman, D. (2020). College and university chief academic officers. Washington, DC: Inside Higher Ed and Gallup.

López-Pastor, V., & Sicilia-Camacho, A. (2017). Formative and shared assessment in higher education. Lessons learned and challenges for the future. Assessment & Evaluation in Higher Education, 42(1), 77-97.

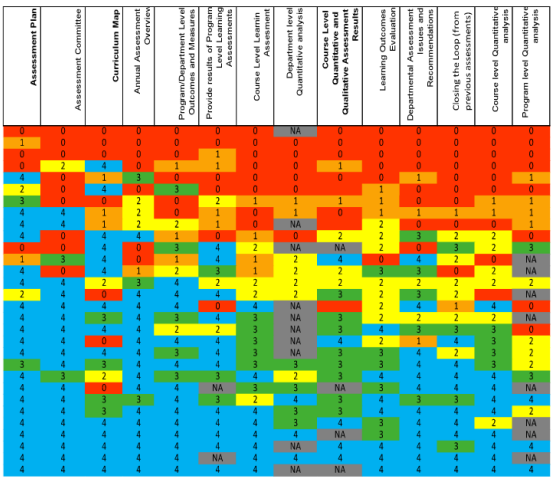
Maki, P. L. (2017). Real-time student assessment: Meeting the imperative for improved time to degree, closing the opportunity gap, and assuring student competencies for 21st-century needs. Sterling, VA: Stylus Publishing

Cookie Cutter Student Learning Reports

A 2019 Council for Higher Education Assessment (CHEA) report on student learning summarized a survey of accreditors:

> In expressing what concerns them most about the future of student learning outcomes and accreditation, institutional and programmatic accreditors were nearly unanimous in their concern about the trend toward standardized "cookie cutter" definitions and approaches as well as the use of "blunt" measures without regard for the rich heterogeneity of the missions, cultures and student populations of institutions and programs. Several institutional accreditors expressed concerns that such a shift will reduce access and stifle innovation in higher education. Programmatic accreditors also raised concerns that overemphasis on rigid or prescriptive student achievement requirements, which may not actually reflect learning, will ultimately diminish attention to other meaningful aspects of students' education experiences.

The reports are sometimes described as "checkbox," which is more descriptive. To see what the checkboxes look like, and what an assessment office's job often is, see Gaudino-Goering, E. (2021). In the figures (see example below), one row represents one program. There will be dozens of programs at a small college and hundreds at a large university.



Management of checkboxes is not a rewarding job, as noted in a survey of assessment staff summarized in "The Assessment Profession in Higher Education: A Snapshot of Perceptions, Roles, and Activities."

> The project manager role captures the logistical/coordination aspects of assessment work. With a few exceptions, this role was most evident among the assessment activities that respondents (both assessment professionals and faculty) liked least about their work. Terms such as "tedious," "cumbersome," or "time consuming" characterized these responses, along with frustrations with data-collection infrastructure or software applications.

Examples of to-do lists for reports can easily be found in presentations on assessment or shared through the ASSESS email list. An example is reproduced below from a presentation slide posted there from Northern Illinois University.

Degree Program: _____ Date: _____
Degree: _____ Reviewer: _____
As you review the assessment plan and status report, note the extent to which you observe each criteria.
N = Not at all, S = Somewhat, and G = Great extent

N	S	G	**Student Learning Outcomes**
			Outcomes are student learning focused
			Outcomes are observable and measurable using verbs (e.g., explain, perform, collaborate)
			A sufficient number of outcomes to add depth/breadth (>3), but not so many to become cumbersome/intrusive (<12); 6-8 tends to be manageable (unless mandated)
			Outcomes differentiate program from similar programs (across degrees and programs)
			Outcomes communicate what successful performance looks like (i.e., performance criteria are implied).
			Outcomes imply higher levels of learning (e.g., analyzing, evaluating, creating)
			Outcomes are aligned with the broader university context (e.g., University Mission, Baccalaureate Goals, College and Unit Mission)
			Outcomes are understandable by a wide range of audiences (e.g., students, faculty, parents, employers, and accreditors)
N	**S**	**G**	**Curriculum Map**
			All Student Learning Outcomes are addressed in a curriculum map
			Multiple courses address each Student Learning Outcome, with higher-level courses addressing SLOs at increasing levels of proficiency

Holistic View: *Incomplete* *Basic* *Proficient* *Advanced*
Observations, comments, and suggestions:

References:

Council for Higher Education Accreditation (2019). Accreditation and Student Success Outcomes: Perspectives from Accrediting Organizations, CHEA/CIQG Publication Series [link]

Gaudino-Goering, E. (2021, February). Using a heat map to visualize academic assessment across the college. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. [link]

Grades as Indirect Evidence

For historical reasons, course grades have been proscribed as primary data for assessing learning, which necessitates the creation of a secondary grading system (see *Cookie Cutter Student Learning Reports*). There are various reasons given to support this ban, summarized by the deprecation of grades as "indirect" evidence. While this does not ban grade data from being used, it relegates it to a secondary status, so that assessment offices have little incentive to analyze grade data for reports on student learning. This largely severs the connection between student learning work and student success research, since grades and course completions are predictors of retention and graduation. It also disconnects learning assessment work from Satisfactory Academic Progress (SAP), a federal rule relating to financial aid eligibility[2]. Good grades are required to progress academically, and grade averages predict writing ability[3], critical thinking test scores[4], graduation rates[5], career outcomes[6]. Grading styles are important to learning[7] and have been shown to be improvable with feedback[8]. Grades are the measures of learning signaled to third parties on transcripts.

In 2019, the Council for Higher Education Accreditation (CHEA) surveyed accreditors on student learning assessment, with a response rate of sixty-four participants, of whom thirteen were institutional

[2] https://ecfr.federalregister.gov/current/title-34/subtitle-B/chapter-VI/part-668/subpart-C/section-668.34

[3] Cite Elliot et all

[4] (1) Kaniuka, T., & Wynne, M. (2019). Exploring the Relationship between the Collegiate Learning Assessment, Student Learning Activities, and Study Behaviors: Implications for Colleges and Universities. *Education Research International*, 2019. (2) Zahner, D., Ramsaran, L. M., & Steedle, J. T. (2012). Comparing alternatives in the prediction of college success. In *Annual Meeting of the American Educational Research Association*, Vancouver, Canada.

[5] The link between grades and staying in school is found even in grade school, e.g., Bowers, A. J. (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts. *The Journal of Educational Research*, 103(3), 191-207. High school grades predict college graduation, often better than standardized tests, e.g. Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., ... & Duckworth, A. L. (2019). Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *American Educational Research Journal*, 56(6), 2077-2115. Since colleges require good grades for academic progress in order to continue, grades necessarily predict graduation in college, but the relationship is closer than that suggests, e.g. Tatar, A. E., & Düştegör, D. (2020). Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average?. *Applied Sciences*, 10(14), 4967.

[6] One study found a correlation of .20 between GPA and earnings, c.f. Roth, P. L., & Clarke, R. L. (1998). Meta-analyzing the relation between grades and salary. *Journal of Vocational Behavior*, 53(3), 386-400.

[7] This paper finds a connection between course difficulty and longitudinal learning (value-added). Insler, M., McQuoid, A. F., Rahman, A., & Smith, K. A. (2021). Fear and Loathing in the Classroom: Why Does Teacher Quality Matter?. *IZA Institute of Labor Economics.* DP 14036.

[8] This paper uses a randomized trial to show that grading consistency can be improved with feedback. Millet, I. (2010). Improving Grading Consistency Through Grade Lift Reporting. *Practical Assessment Research and Evaluation*, 15(4), 1–8.

accreditors. Only five of them (including one of the institutional accreditors) cited course grades as frequently used evidence of student learning, and two of the institutional accreditors reported that they explicitly discourage or ban grades as evidence.

[Two pages of excerpts and comments on accreditor documentation is omitted here. They give examples where peer reviewers refer to grades as "indirect" evidence.]

The concept of indirect versus direct evidence was disputed in the academic (educational measurement) literature by the validity theorist Samuel Messick (1994).

> The portrayal of performance assessments as authentic and direct has all the earmarks of a validity claim but with little or no evidential grounding. That is, if authenticity is important to consider when evaluating the consequences of assessment for student achievement, it constitutes a tacit validity standard, as does the closely related concept of directness of assessment. We need to address what the labels authentic and direct might mean in validity terms. We also need to determine what kinds of evidence might legitimize both their use as validity standards and their nefarious implication that other forms of assessment are not only indirect, but inauthentic. (p. 14)

Indeed, "direct assessment" does not appear in the index of the standard reference Educational Measurement from the National Council on Measurement in Education (Brennan, 2006).

References:

Brennan, R. L., & National Council on Measurement in Education. (2006). *Educational measurement*. Praeger Publishers.

Council for Higher Education Accreditation (2019). Accreditation and Student Success Outcomes: Perspectives from Accrediting Organizations, CHEA/CIQG Publication Series.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(3), 13-23.